Modeling Random Vectors Using Chessboard Distributions

Soumyadip Ghosh and Shane G. Henderson

Department of Operations Research and Industrial Engineering

Cornell University

Ithaca, NY 14853, U.S.A.

March 2, 2004

Abstract

We review chessboard distributions as a tool for modeling random vectors with given marginal distributions and covariance matrix and develop new results on their modeling power. Random vector samples can be rapidly generated from a chessboard distribution after a suitable one is constructed.

1 Introduction

Chessboard distributions are a special class of distributions that can be used to model partially specified random vectors. Typically one specifies the (one dimensional) marginal distributions and some measure of dependence that is usually the (Spearman) rank correlation matrix or the (Pearson) product moment correlation matrix. We review the key ideas behind chessboard distributions and develop new results on their modeling power.

Our motivation for studying chessboard distributions comes from a need in stochastic simulation for an easily applied class of distributions that can capture a range of features of a desired distribution. Indeed, researchers in a variety of fields have sought such a class. See, for example, Devroye (1986), Johnson (1987); refer Ghosh and Henderson (2002) for a survey of these efforts.

In this paper we focus on specifications of marginal distributions and correlations, but one can consider other features such as joint probabilities of certain regions and so forth (Ghosh and Henderson

2001). We believe that for non-Gaussian marginals, it is more appropriate to use rank covariance than product-moment covariance. Recall that the product-moment covariance between two random variables X and Y with distribution functions F and G respectively is given by

$$EXY - EXEY,$$

and the rank covariance is given by

$$EF(X)G(Y) - EF(X)EG(Y).$$

The product moment covariance is well-defined when X and Y have finite second moment, while the rank covariance is always defined. In the case where F and G are continuous, $F(X)$ and $G(Y)$ are uniformly distributed. Hence, one can reduce a study of rank covariance of random vectors with arbitrary continuous distributions to one of product moment covariance of uniform random variables on $(0, 1]$. (We adopt the convention of open intervals on the left and closed on the right. The choice is immaterial for absolutely continuous distributions.)

We therefore focus on the case of generating a random vector with uniform marginals and a desired product moment covariance matrix. (The distribution function of a random vector with uniform marginals on $(0, 1]$ is known as a *copula*. The term was coined in Sklar (1959), and Nelsen (1999) is a useful recent reference.) Many users are not familiar with rank correlation, preferring to work with product moment correlation, so we also provide results for the case of non-uniform marginals and product moment correlation.

We describe chessboard distributions as a subclass of a new class of distributions that we call *replicated copulas*. Our interest in replicated copulas lies primarily in their use in furthering our understanding of chessboard distributons, but they are interesting in their own right. After reviewing chessboard distributions and their relationship to replicated copulas we provide some new results. Specifically, we shed further light on the class of distributions that cannot be exactly matched by chessboard distributions (Theorem 6), we describe the set of covariance matrices that can be matched by a chessboard distribution (Theorem 5), and we extend results known for the uniform marginal case to more general marginals; see Section 5. We also extend some known results for chessboard distributions to replicated copulas.

The remainder of this paper is organized as follows. In Section 2 we review some high-level facts about chessboard distributions and introduce replicated copulas. Then, in Section 3 we describe a linear-programming method for constructing chessboard distributions, and replicating copulas in general.

Section 4 contains what we view as the main results of the paper, and Section 5 discusses the case of non-uniform marginal distributions.

2 Chessboard Distributions and Replicated Copulas

A chessboard copula is a member of a class of copulas with a special structure. We call this the family of *replicated* copulas. The structure of a replicated copula for X is easily described. For notational convenience we confine our discussion to the three-dimensional case. The d dimensional case is similar. We divide $(0,1]^3$ into a large grid of rectangular regions (cells) with sides parallel to the coordinate axes. Let $n \geq 1$ be an integral parameter that determines the level of division that is performed. The range $(0,1]$ of the ith variable is divided into n equal-length sub-intervals by the set of points $y_{i,k} = \frac{k}{n},\ k = 0, \ldots, n$. Denote the cells as $C(j_1, j_2, j_3)$, indexed by $j_1, j_2, j_3 = 1, \ldots, n$. Within each cell $C(j_1, j_2, j_3)$ the joint distribution follows an appropriately scaled and translated version of a copula $\mathcal{C}(j_1, j_2, j_3)$. We call this copula the *base* copula. Suppose $Z = (Z_1, Z_2, Z_3)$ is a random vector distributed as $\mathcal{C}(j_1, j_2, j_3)$. Then, conditional on being in cell $C(j_1, j_2, j_3)$, the joint distribution of X can be obtained from Z as

$$X_1 = \frac{Z_1}{n} + y_{1,j_1-1},\ \forall j_1 = 1, \ldots, n, \quad \text{and similarly for } X_2, X_3. \tag{1}$$

The base copulas could be the same or could vary over different cells. We limit ourselves to replicated copulas that have the same base copula in each cell. So, in essence we divide the region $(0,1]^3$ into non-overlapping cells and replicate a given copula within the cells. Theorem 1 in Section 3 shows that a function created by this replication operation is a valid copula.

Johnson and Kotz (2004) study similar replicated bivariate distributions, which they term *cloned* distributions. They also assume the mass assigned to each cell to be the same, while we allow it to vary. To be more precise, let $q(j_1, j_2, j_3) = P(X \in C(j_1, j_2, j_3))$ be the mass assigned to the copula replicated at cell $C(j_1, j_2, j_3)$. Our copula construction technique picks values for $q(j_1, j_2, j_3)$ so that the desired covariance matrix is achieved, subject to the constraint that the constructed copula is a valid joint distribution function. This is done by solving a linear program formulated with the $q(j_1, j_2, j_3)$s as variables. The technique concludes either by finding $q(j_1, j_2, j_3)$s that give a joint distribution for X with the desired properties, or by determining that no joint distribution can be constructed for X with these properties.

Chessboard copulas, as introduced in Ghosh and Henderson (2002), are replicated copula where the

base copula is that of independent uniform random variables. Chessboard distributions are essentially the "piecewise-uniform copulae" that Mackenzie (1994) developed. Mackenzie (1994) identifies chessboard copulas with maximum entropy that match a given covariance matrix, assuming that the covariance matrix can be matched. We do not make this assumption, and develop this family partly to provide a procedure to check whether given covariance matrices can be matched. The product-form copula has a density such that each component is independent of the other, and hence its replication in cells makes the components of X conditionally independent (conditional on being in the cell) with marginal distributions given by the uniform distribution restricted to the cell. This special form has an advantage in that it leads to a simple scheme for generating samples from the chessboard copula.

There are many possible methods for generating random vectors with replicated copulas. The methods vary in terms of their time and storage requirements for setup, and for generating random vectors once the setup is complete. What follows is a generic approach that requires a moderate amount of time and storage for setup, but once the setup is complete requires very little time to generate random vectors. Let d denote the dimension of the random vector X with marginal distribution functions $F_1, \ldots, F_d$. Suppose that $q(\cdot, \ldots, \cdot)$ and $\mathcal{C}$ together represent the replicated distribution constructed for X. The algorithm is as follows.

1. Generate the indices $(j_1, \ldots, j_d)$ of the cell containing X from the probabilities $q(\cdot, \ldots, \cdot)$.
2. Generate X from its conditional distribution given that it lies in the cell $C(j_1, \ldots, j_d)$: an appropriately scaled version of $\mathcal{C}$.

The first step can be performed efficiently using, for example, the alias method. The alias method, developed by Walker (1977) and discussed in detail in Law and Kelton (2000), can generate the appropriate cell in constant time, and requires $O(m)$ storage and $O(m)$ setup time, where m is the number of positive $q(j_1, j_2, \ldots, j_d)$ values. If $q(\cdot, \ldots, \cdot)$ is an extreme-point solution to one of the linear programs developed in Section 3, then there are on the order of nd strictly positive cell probabilities. This follows from a standard result in linear programming that any extreme point solution to a system of m linear equalities in nonnegative variables has at most m strictly positive values. The exact number of positive values depends on the number of equality constraints in the LP and the degree to which the extreme-point solution is degenerate. (A degenerate extreme-point solution is one with less than m strictly positive values.)

The fact that $m = O(nd)$ is relatively small can be viewed as an advantage with respect to variate

generation since it reduces the setup time required to implement the alias method. However, it can also be viewed as a disadvantage in terms of modeling power. For a given dimension d and discretization level n there are n^d cells. Of these, $O(nd)$ receive strictly positive probabilities $q(j_1, \ldots, j_d)$. So as the dimension d increases, the fraction of cells receiving positive probabilities is vanishingly small. This means that the set of values that the random vector X can assume is somewhat limited.

Mackenzie (1994) avoids this problem by maximizing the entropy of the discrete distribution $q(\cdot, \ldots, \cdot)$. In this case, all of the cells receive positive probability. However, the problem of maximizing the entropy of q subject to linear constraints is a convex optimization problem that is more difficult to solve than the LPs discussed in this paper.

Suppose cell $C(j_1, \ldots, j_d)$ is chosen in Step 1 above. Conditional on X lying in this cell, the components $X_1, \ldots, X_d$ of X are distributed according to a transformed version of $\mathcal{C}$. Suppose $Z = (Z_1, \ldots, Z_d)$ is a random vector distributed as $\mathcal{C}$. A sample of X can be obtained by first sampling a Z and then transforming Z as in (1). Thus, for instance, if $\mathcal{C}$ is the product-form coupla, as in Ghosh and Henderson (2002), then in Step 2, we can independently generate each component from its respective conditional (marginal) distribution. The efficiency of this step clearly depends on the form of $\mathcal{C}$. The product-form base copula requires d independent uniform random variables to generate a sample of Z. On the other hand, if the base copula were the *maximally-correlated* copula, with a form such that $Z_1 = Z_2 = \ldots = Z_d$, then only one uniform random variable need be generated to get a sample of Z.

3 Constructing a Chessboard Copula

In this section we focus on chessboard copulas, i.e., replicated copulas using the uniform distribution on the unit hypercube as the replication copula. However, we explain how certain results extend to replicated copulas. For notational simplicity we confine our attention to the three-dimensional case. The extension to higher dimensions is straightforward.

Our goal is to construct the density of a random vector X with uniform marginals on (0,1] and product-moment covariance matrix $\Sigma = (\Sigma_{ij} : 1 \leq i, j \leq 3)$.

Let $n \geq 1$ be an integral parameter that determines the level of discretization that will be performed. Divide $(0, 1]^3$ into a grid of n^3 rectangular regions (cells) with sides parallel to the coordinate axes. Let $\{y_{i,k} = \frac{k}{n}, k = 0, \ldots, n\}$ be the set of points that divide the range (0,1] of the ith variable into n equal

length sub-intervals. We define the $C(j_1, j_2, j_3)$th cell to be the set

$$\{(x_1, x_2, x_3) : y_{i,j_i-1} < x_j \leq y_{i,j_i},\ i = 1, 2, 3\},$$

for $1 \leq j_1, j_2, j_3 \leq n$. The density f of X is piecewise constant, taking the value $q(j_1, j_2, j_3)n^3$ in the cell $C(j_1, j_2, j_3)$, so that

$$P(X \in C(j_1, j_2, j_3)) = q(j_1, j_2, j_3). \tag{2}$$

For the general case of replicated copulas, X may not necessarily have a density, but (2) still holds.

To ensure that the $q(j_1, j_2, j_3)$ form a set of mixing coefficients and the density f of X has uniform marginals we require that

$$\begin{aligned}
\sum_{j_2,j_3=1}^{n} q(j_1, j_2, j_3) &= P(X_1 \in (y_{1,j_1-1},\, y_{1,j_1}]) = \frac{1}{n}, \quad \forall j_1 = 1, \ldots, n, \\
\sum_{j_1,j_3=1}^{n} q(j_1, j_2, j_3) &= P(X_2 \in (y_{2,j_2-1},\, y_{2,j_2}]) = \frac{1}{n}, \quad \forall j_2 = 1, \ldots, n, \\
\sum_{j_1,j_2=1}^{n} q(j_1, j_2, j_3) &= P(X_3 \in (y_{3,j_3-1},\, y_{3,j_3}]) = \frac{1}{n}, \quad \forall j_3 = 1, \ldots, n, \\
q(j_1, j_2, j_3) &\geq 0 \quad \forall j_1, j_2, j_3 = 1, \ldots, n.
\end{aligned} \tag{3}$$

Assuming (3) holds, it then follows that X has the desired uniform marginals.

Theorem 1 *If the distribution of X is a replicated copula with cell probabilities q satisfying the constraints (3), then X has uniform marginals.*

A proof of this result for the chessboard copula case can be found in Ghosh and Henderson (2002). We present this slightly generalized version because it is useful in understanding the nature of replicated distributions.

Proof: Let the marginal distribution function of X_1 be denoted by $F_1(\cdot)$. We have to show that $F_1(x) = x$ for $x \in (0, 1]$. Let Z represent a random vector corresponding to the base copula $\mathcal{C}$. The components of X and Z are related as in (1). For any $x \in (y_{1,i-1},\ y_{1,i}]$, we have that

$$\begin{aligned}
F_1(x) &= \sum_{j_1 \leq i-1} \sum_{j_2,j_3=1}^{n} q(j_1, j_2, j_3) + \sum_{j_2,j_3=1}^{n} P(y_{1,i-1} < X_1 \leq x | X \in C(j_1, j_2, j_3)) \cdot q(j_1, j_2, j_3) \\
&= \frac{i-1}{n} + \sum_{j_2,j_3=1}^{n} P(0 < Z_1 \leq n(x - y_{1,i-1})) \cdot q(j_1, j_2, j_3) \\
&= \frac{i-1}{n} + \sum_{j_2,j_3=1}^{n} n(x - \frac{i-1}{n}) q(j_1, j_2, j_3)
\end{aligned}$$

$$= \quad \frac{i-1}{n} + x - \frac{i-1}{n} = x$$

as required. The first equation follows by conditioning on the cell in which the random vector lies, and the second is obtained from (3) and the transformation that relates X with Z. The third equation uses the fact that Z_1 is uniformly distributed and the final equation again uses (3). A similar result holds for X_2 and X_3, and so the joint distribution has uniform marginals. □

Recall that our goal is to match the correlation matrix Σ. We do this using a linear program. If $\Sigma^X_{ij} = \mathrm{Cov}(X_i, X_j)$ gives the covariances of the random vector X with density f, then we want to minimize the distance $r(\Sigma^X, \Sigma)$ between Σ^X and Σ, where

$$r(\Sigma^X, \Sigma) = \sum_{1 \le i < j \le 3} \left| \Sigma^X_{ij} - \Sigma_{ij} \right|.$$

Now, X has uniform marginals so $EX_i = 1/2$ for $i = 1, 2, 3$. Also, by conditioning on the cell containing X we see that

$$\begin{aligned} EX_1X_2 &= \sum_{j_1, j_2, j_3} q(j_1, j_2, j_3) E[X_1X_2 | X \in C(j_1, j_2, j_3)] \\ &= \sum_{j_1, j_2, j_3} q(j_1, j_2, j_3)\ \mu_{1,j_1}\ \mu_{2,j_2}, \end{aligned} \tag{4}$$

where

$$\mu_{\ell,i} = E[X_\ell | X \in (y_{\ell,i-1}, y_{\ell,i}]] = \frac{2i-1}{2n}$$

is the conditional mean of X_ℓ given that it lies in the ith subinterval. In (4) we used the conditional independence of the components of X given that X lies in one of the cells. In the general case of replicated copulas the first equality remain unchanged, and one again obtains a weighted sum of $q(j_1, j_2, j_3)$s in (4) but with different weights.

It follows that Σ^X_{12} is a linear function of the $q(j_1, j_2, j_3)$'s, as is Σ^X_{13} and Σ^X_{23}. Using a standard trick in linear programming, we can represent $|\Sigma^X_{12} - \Sigma_{12}|$ and the other terms in $r(\Sigma^X, \Sigma)$ in a linear fashion as follows.

Define Z^+_{ij} and Z^-_{ij} to be the positive and negative parts of the difference $\Sigma^X_{ij} - \Sigma_{ij}$, i.e.,

$$Z^+_{ij} = (\Sigma^X_{ij} - \Sigma_{ij})^+ = \max\{\Sigma^X_{ij} - \Sigma_{ij}, 0\}, \text{ and } (\Sigma^X_{ij} - \Sigma_{ij})^- = -\min\{\Sigma^X_{ij} - \Sigma_{ij}, 0\}.$$

We can now attempt to match Σ^X to X using the LP

$$\min \qquad \sum_{i=1}^{2} \sum_{j=i+1}^{3} (Z^+_{ij} + Z^-_{ij}) \tag{5}$$

$$\text{subject to} \qquad \Sigma^X_{ij} - \Sigma_{ij} = Z^+_{ij} - Z^-_{ij},\ i = 1 \text{ to } 2 \text{ and } j = i+1 \text{ to } 3$$

$$Z^+_{ij} \geq 0, Z^-_{ij} \geq 0, \text{ together with constraints (3).}$$

This LP is always feasible since we can set $q(j_1, j_2, j_3) = n^{-3}$ for all j_1, j_2, j_3. Also, the objective function of the LP is bounded below by 0, so an optimal solution exists.

If the optimal objective value for the LP is 0, then the solution gives a distribution with the desired marginals and covariance structure, i.e., $\Sigma^X = \Sigma$. This provides the desired construction. (This observation goes through unchanged for replicated copulas.)

Recall that we also want a test that can establish that certain matrices Σ cannot be matched. To this end we develop bounds on the Z^+_{ij} and Z^-_{ij} variables. These additional bounds are important, because if they cannot be satisfied by a feasible solution to the LP then a random vector with the given covariance matrix and uniform marginals does not exist, as discussed further in Section 4.

The bounds are developed by assuming that a random vector X with uniform marginals and covariance matrix Σ exists, and modifying the distribution of X to that of a random vector $\tilde{X}$ that has a chessboard distribution. The modification consists of keeping the total mass within each cell constant, but making the conditional distribution within the cell uniform. (In the case of replicated copulas the conditional distribution within the cell is a scaled version of the base copula.) The distribution of $\tilde{X}$ then gives a feasible solution to the LP (minus the bounds). We can bound the change in the covariances resulting from this modification of the distribution.

Let

$$\tilde{q}(j_1, j_2, j_3) = P(X \in C(j_1, j_2, j_3)) = P(\tilde{X} \in C(j_1, j_2, j_3)).$$

Observe that

$$\begin{aligned} \text{Cov}(\tilde{X}_1, \tilde{X}_2) - \Sigma_{12} &= E\tilde{X}_1\tilde{X}_2 - EX_1X_2 \\ &= \sum_{j_1,j_2,j_3=1}^{n} (\mu_{1,j_1}\mu_{2,j_2} - E[X_1X_2|X \in C(j_1, j_2, j_3)])\ \tilde{q}(j_1, j_2, j_3). \end{aligned} \tag{6}$$

But

$$y_{1,j_1-1}\ y_{2,j_2-1} \leq E[X_1X_2|X \in C(j_1, j_2, j_3)] \leq y_{1,j_1}\ y_{2,j_2}. \tag{7}$$

Combining (6) with (7) we see that

$$\text{Cov}(\tilde{X}_1, \tilde{X}_2) - \Sigma_{12} \leq \sum_{j_1,j_2,j_3=1}^{n} \tilde{q}(j_1, j_2, j_3)(\mu_{1,j_1}\ \mu_{2,j_2} - y_{1,j_1-1}\ y_{2,j_2-1}) \text{ and} \tag{8}$$

$$\text{Cov}(\tilde{X}_1, \tilde{X}_2) - \Sigma_{12} \geq \sum_{j_1,j_2,j_3=1}^{n} \tilde{q}(j_1, j_2, j_3)(\mu_{1,j_1}\ \mu_{2,j_2} - y_{1,j_1}\ y_{2,j_2}). \tag{9}$$

Equation (8) gives an upper bound on Z_{12}^{+}, and (9) gives an upper bound on Z_{12}^{-}. Similar bounds may be obtained for the other covariances. After substituting in the explicit expressions for $y_{i,k}$ and $\mu_{i,k}$, these bounds simplify to

$$Z_{ij}^{+} \leq \frac{1}{2n} - \frac{1}{4n^2} \quad \text{and} \quad Z_{ij}^{-} \leq \frac{1}{2n} + \frac{1}{4n^2} \quad 1 \leq i < j \leq 3. \tag{10}$$

The linear program (5) is henceforth assumed to include the bounds (10).

A similar approach can be adopted for replicated copulas. In this case, one again obtains bounds like (8) and (9), but with different values replacing $\mu_{1,j_1}\ \mu_{2,j_2}$. One then gets bounds analogous to (10) that are again of the order n^{-1}.

4 Modeling Power

We have introduced linear programs for constructing chessboard distributions with uniform marginals and covariance matrix. But how effective are these methods? In this section we summarize key results from Ghosh and Henderson (2002) without proof. In giving these results, we allow the random vector to have arbitrary, but finite, dimension $d > 1$. We restrict attention to uniform marginals, but the results extend to other marginals with densities and finite variances as explained in Section 5. We again focus on the chessboard case, and point out generalizations to the replicated copula case where appropriate.

Definition 1 *We say that a covariance matrix Σ is* feasible *for a given set of marginal distributions if a random vector with the given marginals and covariance matrix exists.*

Theorem 2 *A covariance matrix is infeasible for the given marginals if, and only if, the chessboard LP is infeasible for some $n \geq 1$.*

This result establishes that if one of the LPs is infeasible for *any* discretization level n, then the proposed covariance matrix is infeasible. Furthermore, the theorem establishes that if a covariance matrix is infeasible, then one will eventually discover this by solving an LP with n sufficiently large. To our knowledge, this is the first example of a tight characterization of infeasible covariance matrices for random vectors of dimension $d \geq 3$.

The same result holds for replicated copulas, where the LP (5) is supplemented with the appropriate bounds.

Of course, we are more interested in a positive result. Given the sharp characterization in Theorem 2, it would be nice if chessboard distributions could *exactly* match any arbitrary feasible covariance matrix. Unfortunately, this is not the case, as the following example shows.

Example 1 *Suppose that Z is a $2-$dimensional random vector with uniformly distributed components $Z_1 = Z_2$ on $(0, 1]$, so that the*

$$Cov(Z_1, Z_2) = Var(Z_1) = 1/12.$$

For a bivariate chessboard random vector X of a given size n, the covariance between X_1 and X_2 is maximized by concentrating all mass on the cells (i, i), and so $q(i, i) = n^{-1}$ for $1 \leq i \leq n$. In that case, we have that

$$Cov(X_1, X_2) = \frac{1}{12} - \frac{1}{12n^2}.$$

Therefore, $Cov(X_1, X_2) < 1/12 = Cov(Z_1, Z_2)$ for all finite n.

This example shows that chessboard distributions cannot exactly match all feasible covariance matrices. Notice though, that the error in the covariance matrix can be made arbitrarily small. In fact it is possible to show that chessboard distributions can arbitrarily-closely approximate any feasible covariance matrix.

Theorem 3 *Suppose that Σ is feasible. Then for all $\epsilon > 0$, there exists a chessboard distribution with covariance matrix Λ with the property that $r(\Sigma, \Lambda) < \epsilon$.*

Theorem 3 also holds for replicated copulas, as can be shown using essentially exactly the same proof as for chessboard distributions.

Not only can chessboard distributions closely approximate any feasible covariance matrix, but they can *exactly* match "almost all" feasible covariance matrices. To formulate and state this result precisely, we need some more terminology and a definition.

With an abuse of notation, we can view a $d \times d$ covariance matrix as an element of $d(d-1)/2$ dimensional space. This follows because there are $d(d-1)/2$ elements above the diagonal, the matrix is symmetric, and the diagonals are determined by the marginal distributions. Let Ω denote the set of feasible covariance matrices. We view this set as a subset of $d(d-1)/2$ dimensional Euclidean space. Ghosh and Henderson (2002) prove the following two results.

Proposition 4 *The set Ω is nonempty, convex, closed and full-dimensional.*

Let A° and ∂A denote the interior and boundary of the set A respectively.

Theorem 5 *There is a chessboard distribution with covariance matrix Σ if, and only if, $\Sigma \in \Omega^\circ$.*

The "if" part of Theorem 5 remains true for replicated copulas, but it is also possible for replicated copulas to achieve some points on the boundary of Ω. For instance, continuing Example 1, suppose that the base copula corresponds to a perfectly correlated pair of uniform random variables. Then one can achieve a correlation of 1 with $n = 1$.

Proposition 4 establishes that the set Ω has a non-zero finite Lebesgue measure, while $\partial\Omega$ is a zero Lebesgue measure set. It follows from Theorem 5 that chessboard disributions can be constructed for almost any (in a Lebesgue measure sense) feasible covariance matrix from Ω. Given a feasible covariance matrix, the procedure to determine a corresponding chessboard distribution is then straightforward: one solves the augmented LP based on (5) for a chosen level of discretization n and if the optimal objective value is greater than 0, the parameter n is increased successively till the objective value drops to 0 or an acceptable value.

Theorem 5 establishes that no covariance matrix in $\partial\Omega$ can be matched by chessboard distributions. We can prove a slightly stronger result regarding any distribution F that has a covariance matrix $\Sigma \in \partial\Omega$ (and uniform marginals). The distribution F can be decomposed into a singular part F_s and an absolutely continuous part F_{ac} with respect to Lebesgue measure restricted to $(0, 1]^3$ (the *Lebesgue Decomposition*). Thus,

$$F = F_{ac} + F_s.$$

Moreover, the absolutely continuous part has a density f_{ac} in the sense of the *Radon-Nikodym derivative* of F_{ac}.

Theorem 6 *There cannot exist an open set G such that*

$$f_{ac}(x) \geq \phi > 0 \quad \forall x \in G, \text{ except on a subset of Lebesgue measure} 0. \tag{11}$$

Proof: For notational ease we give a proof in the 3-dimensional case. The general case is virtually identical. Suppose such a G exists. We can reassign f_{ac} to have value ϕ over any subset of measure zero where the f_{ac} cannot be bounded away from 0, without changing the function F_{ac}. Thus, we assume that f_{ac} is bounded away from 0 by atleast ϕ over all $x \in G$. We can choose an open ball $B(x, \epsilon)$ within

G and an open cubical region C with sides aligned to the axes within $B(x, \epsilon)$ such that the interior of C is non-empty. Split f_{ac} into two parts f_C and $f_{\bar{C}}$ defined as:

$$f_C(x) = \begin{cases} \phi & x \in C \\ 0 & \text{elsewhere} \end{cases} \quad \text{and} \quad f_{\bar{C}}(x) = \begin{cases} f_{ac}(x) - \phi & x \in C \\ f_{ac} & \text{elsewhere} \end{cases}.$$

Let u and v be the endpoints that define C so that

$$C = \{(x, y, z) \in (0, 1]^3 : u_1 < x \leq v_1, u_2 < y \leq v_2, u_3 < z \leq v_3\}.$$

Divide the cell C into 4 (equal size) subcells,

$$\begin{aligned} C_{ab} &= \{(x, y, z) \in C : u_1 + (a-1)\frac{v_1 - u_1}{2} < x \leq u_1 + a\frac{v_1 - u_1}{2}, \\ &\quad u_2 + (b-1)\frac{v_2 - u_2}{2} < x \leq u_2 + b\frac{v_2 - u_2}{2}\} \end{aligned}$$

for $1 \leq a, b \leq 2$.

Define a new distribution H from F by leaving F unchanged in all of $(0, 1]^3$ except over C. Within the cell C, assign a density h_C of 2ϕ to each of the cells C_{11}, and C_{22}, and set h_C to be 0 in the cells C_{ab} for $a \neq b$. Then it is straightforward to show that H has uniform marginals, that the $(1, 2)$th covariance is strictly increased, and that the other covariances remain unchanged. A similar argument increasing the density in the cells C_{ab} with $a \neq b$ shows that the covariance can be strictly decreased.

Convexity of Ω then implies that Σ must lie in the interior Ω° which is a contradiction, and the proof is complete.□

It is conceivable that the support of distributions that match matrices from $\partial\Omega$ could consist of sets of zero Lebesgue measure in $\mathbb{R}^d$ or exotic sets like Cantor sets with no interior but non-zero measure. Generating random vectors with such distributions could thus prove difficult.

Theorem 6 also tells us that if one uses a base copula with sets as described by (11) in its support, then one cannot construct replicated copulas to exactly match covariance matrices from $\partial\Omega$.

In summary then, chessboard distributions

- can detect if a given matrix is infeasible,
- can arbitrarily closely approximate any feasible covariance matrix,
- can exactly match any feasible covariance matrix in the interior of the set of feasible covariance matrices, but

- cannot exactly match any covariance matrix on the boundary of the set of feasible covariance matrices.

In Section 2, we propose a method to generate from chessboard distributions. We have posited that once the chessboard distribution is set up, generation should be fast. However, for the method to be viable, one should be able to set up the distribution in a reasonable amount of time. Critical to the distribution determination step is the time it takes to obtain a solution for the linear programs based on (5). Efficient algorithms are available to solve these linear programs, and are known to be theoretically solvable in time which is a polynomial function of the size (in binary representation) of the problem. The setup time thus depends on the size n of the discretization that is used. We now turn to the question of how large n needs to be to match a given covariance matrix in Ω°.

Let

$$S(\epsilon) \stackrel{\triangle}{=} \{\Sigma \in \Omega : (1+\epsilon)\Sigma \in \Omega\} = \{A \in \mathbb{R}^{d(d-1)/2} : A = (1+\epsilon)^{-1}\Sigma, \text{ for some } \Sigma \in \Omega\}$$

be the set that represents a contour set of Ω indexed by ϵ. This is essentially the set Ω scaled down by a factor $(1+\epsilon)^{-1}$. Proposition 4 gives us that the set $S(\epsilon)$ is a convex, closed, bounded and full dimensional subset of Ω.

Theorem 7 *Let $\Omega^n \subseteq \Omega$ be the set of covariance matrices that can be matched by chessboards of size n. Then $\Omega^n \subseteq S(\frac{1}{n^2})$.*

Proof: For notational ease we prove the result for $d = 3$. The case $d > 3$ is proved similarly. The underlying idea of the proof is that one can move in every direction from any Σ in Ω^n and remain in Ω, and one can obtain a bound on the distance one can move, which gives the result.

Let $\{q(j_1, j_2, j_3)\}$ represent an LP solution that constructs a chessboard distribution corresponding to covariance matrix Σ. Then

$$\begin{aligned} \Sigma_{12} &= EX_1X_2 - EX_1EX_2 \\ &= \sum_{j_1,j_2,j_3=1}^{n} E[X_1X_2|X \in C(j_1,j_2,j_3)] \cdot q(j_1,j_2,j_3) - \frac{1}{4}. \end{aligned} \tag{12}$$

Let $Z = (Z_1, Z_2, Z_3)$ be a random vector endowed with the base copula being replicated within the cells and $\Sigma^Z \in \Omega$ be its covariance matrix. In our case of a chessboard distribution, Z is a vector of independent uniform random variables and $\Sigma^Z = (0,0,0)$. Let y_{i,j_i}, $i = 1,2,3$, $j_i = 1,\ldots,n$ be the

points that define the grid as in (1). Since $EZ_i = 1/2$, $i = 1, 2, 3$, we see that

$$\begin{aligned} E[X_1 X_2 | X \in C(j_1, j_2, j_3)] &= E\left[\left(\frac{Z_1}{n} + y_{1,j_1-1}\right)\left(\frac{Z_2}{n} + y_{2,j_2-1}\right)\right] \\ &= \frac{EZ_1 Z_2}{n^2} + \frac{EZ_1\, y_{2,j_2-1} + EZ_2\, y_{1,j_1-1}}{n} + y_{1,j_1-1}\, y_{2,j_2-1} \\ &= \frac{EZ_1 Z_2}{n^2} + \frac{y_{2,j_2-1} + y_{1,j_1-1}}{2n} + y_{1,j_1-1}\, y_{2,j_2-1} \\ &= \frac{EZ_1 Z_2}{n^2} + t(j_1, j_2), \end{aligned} \tag{13}$$

where $t(j_1, j_2)$ is a function only of the indices j_1 and j_2.

Suppose now that we replace the product-form copula in each cell of the chessboard distribution with another copula represented by the random vector Z'. The result is still a valid replicated copula because of Theorem 1, and represents the distribution of a random vector X' say. If Σ' is the covariance matrix of X', then

$$\begin{aligned} \Sigma'_{12} &= \sum_{j_1,j_2,j_3=1}^{n} E[X'_1 X'_2 | X' \in C(j_1, j_2, j_3)] \cdot P(X' \in C(j_1, j_2, j_3)) - \frac{1}{4} \\ &= \sum_{j_1,j_2,j_3=1}^{n} \left(\frac{EZ'_1 Z'_2}{n^2} + t(j_1, j_2)\right) \cdot q(j_1, j_2, j_3) - \frac{1}{4}. \end{aligned} \tag{14}$$

Let $\Sigma^{Z'}$ be the covariance matrix of Z'. The net change in covariance due to the replacement operation is, from (12), (13) and (14),

$$\begin{aligned} \Sigma'_{12} - \Sigma_{12} &= \sum_{j_1,j_2,j_3=1}^{n} \frac{1}{n^2}(EZ'_1 Z'_2 - EZ_1 Z_2) \cdot q(j_1, j_2, j_3) \\ &= \frac{1}{n^2}(\Sigma^{Z'}_{12} - \Sigma^{Z}_{12}) \\ &= \frac{1}{n^2}\Sigma^{Z'}_{12}. \end{aligned} \tag{15}$$

We have used the fact that $\Sigma^Z_{12} = 0$, since Z represents the product-form copula. Equation (15) holds for every component of the covariance matrix. Hence,

$$\Sigma' = \Sigma + \frac{1}{n^2}\Sigma^{Z'}.$$

Observe that Z' can have any arbitrary covariance matrix in Ω, including those from the boundary $\partial\Omega$. Thus,

$$\Omega^n \subseteq \bar{S}(n) \stackrel{\triangle}{=} \{\Sigma \in \Omega : \Sigma + \frac{1}{n^2}\Lambda \in \Omega,\ \forall \Lambda \in \Omega\}.$$

If $A \in \bar{S}(n)$, then it follows from the definition of $\bar{S}(n)$ that $A(1 + \frac{1}{n^2}) \in \Omega$. Hence, $A \in S(\frac{1}{n^2})$, which implies that $\bar{S}(n) \subseteq S(\frac{1}{n^2})$. This gives us the result. □

This result shows that chessboard distributions with discretization level n cannot come closer than within a factor $(1 - \frac{1}{n^2})$ of the points on $\partial\Omega$. The result is tight in a certain sense. From Example 1, a chessboard of size n can come to within $\frac{1}{12n^2}$ of achieving a covariance value of $\frac{1}{12}$.

An analogous result holds for the general case of constructing replicated copulas whose base copula corresponds to a covariance matrix from the interior Ω°. Some straightforward modifications to the expressions that lead to (15) show that in this case we can again move in every possible direction and the points within the set Ω^n will again be a similar factor of the order of $(1 + \frac{1}{n^2})^{-1}$ away from the boundary $\partial\Omega$.

5 Chessboards With General Marginals

In this section we generalize the uniform marginals assumption to consider random vectors with marginal distributions that have densities and finite variance. The requirement that the marginal distributions have densities is again for convenience. It allows us to be less stringent with endpoints of intervals than we would otherwise need to be.

The theory developed in the earlier sections of this paper can be translated to this case in a straightforward manner for the most part. Suppose we wish to construct a chessboard distribution for $X = (X_1, X_2, X_3)$ such that each of the marginals has a density f_i and finite variance. Let

$$\{y_{i,j_i} : i = 1, 2, 3,\ j_i = 0, \ldots, n\}$$

be the set of points that divide the range of the ith variable into n sub-intervals. The range could be infinite, in which case we allow the corresponding endpoint to be $\pm\infty$. Let M_i^- and M_i^+ represent the leftmost and rightmost finite points respectively. Thus, if X_i were exponentially distributed, $M_i^- = y_{i,0} = 0, y_{i,n} = \infty$ and $M_i^+ = y_{i,n-1}$. The range can be divided in any manner, as long as the internal mesh becomes dense, i.e.,

$$\lim_{n\to\infty} \sup_{i,j_i} |y^n_{i,j_i} - y^n_{i,j_i-1}| = 0,$$

where the sup excludes infinite endpoints, and $\min_i |M_i^\pm| \to \infty$ as $n \to \infty$. These conditions are satisfied, for example, if we take the points to be of the form

$$-\sqrt{n}, -\sqrt{n} + \frac{2}{\sqrt{n}}, -\sqrt{n} + \frac{4}{\sqrt{n}}, \ldots, \sqrt{n}.$$

The chessboard distribution is defined so that within each cell the components of X are independent, and are distributed according to the desired marginals f_1, f_2, f_3 restricted to the cell $C(j_1, j_2, j_3)$. Let

$p_{i,j} = P(X_i \in (y_{i,j-1}, y_{i,j}])$ be the probability that the ith marginal random variable lies in the jth subinterval. The density $f(x)$ of X evaluated at $x \in C(j_1, j_2, j_3)$ is then given by

$$q(j_1, j_2, j_3) \frac{f_1(x_1)}{p_{1,j_1}} \frac{f_2(x_2)}{p_{2,j_2}} \frac{f_3(x_3)}{p_{3,j_3}}. \tag{16}$$

An argument along the lines of the proof of Theorem 1 shows that if a chessboard distribution is constructed via (16) then it has the correct marginals.

Theorem 8 *If the distribution of X is given by (16), then X has the correct marginal distributions.*

The results given in Section 4 and in Ghosh and Henderson (2002) hinge on the crucial fact that one is able to obtain bounds on the objective function of the LPs (5) that vanish as the discretization parameter $n \to \infty$. These proofs can be modified to work for general marginals with densities and finite variances if a similar vanishing bound can be identified. The technique employed in the uniform marginal distribution case (Section 3) does not carry over to this case since it depends on the support being finite. We present a technique to derive such bounds under the assumption that the marginal distributions have finite variance, which is an assumption that is required anyway to ensure that the covariances are well-defined. This then extends the power of the chessboard construction technique to the general marginals case.

We shall restrict our description to the case where the set of points that form the support of the marginal densities are all symmetric about 0 and with infinite support. We assume this only to keep the notation simple; the method itself is applicable to any distribution with infinite support. The symmetry implies that $|M_i^-| = |M_i^+| \stackrel{\triangle}{=} M_i$. Suppose there exists a random vector X with the prescribed covariance matrix Σ. We again redistribute the probability mass of its distribution within cells (thus keeping the cell probability masses constant) so that the conditional density given a cell is one of independent random variables with the desired marginals. Let $\tilde{X}$ denote a random vector with the redistributed probability mass. We provide a bound on the change in covariance due to this redistribution. We have that

$$E[X_1X_2] = E[X_1X_2\ 1\{\ \{|X_1| \le M_1\} \cap \{|X_2| \le M_2\}\ \}] + E[X_1X_2\ 1\{\ \{|X_1| > M_1\} \cup \{|X_2| > M_2\}\ \}], \tag{17}$$

where $1\{A\}$ is the indicator function taking the value 1 on the event and 0 otherwise. The first term represents the part of the support of X bounded by the rectangle $[-M_1, M_1] \times [-M_2, M_2]$. The change due to the redistribution operation in this part of the second moment of X can be bounded in a fashion similar to that used in Section 3. The last term includes the cells of infinite length. Note that the last

term is

$$\begin{aligned} & E[X_1X_2\ 1\{\ \{|X_1| > M_1\} \cup \{|X_2| > M_2\}\ \}] \\ \leq\ & E[X_1 1\{|X_1| > M_1\}X_2] + E[X_1X_2 1\{|X_2| > M_2\}]. \end{aligned} \tag{18}$$

Let us consider the first term in (18). Since the variances of all components of X are finite, we have, by the Cauchy-Schwarz inequality, that

$$\begin{aligned} |E[X_1 1\{|X_1| > M_1\}X_2]| &\leq E[|X_1 1\{|X_1| > M_1\}|^2\ ]^{1/2} E[X_2^2]^{1/2} \\ &= E[X_1^2\ 1\{|X_1| > M_1\}]^{1/2} E[X_2^2]^{1/2}. \end{aligned} \tag{19}$$

The second term in (19) is a constant that depends on the marginal distribution of X_2. The first term depends on the marginal distribution of X_1 and converges to 0 as $M_1 \to \infty$, due to the finite second moment assumption.

The absolute change in covariance due to the redistribution operation, $|EX_1X_2 - E\tilde{X}_1\tilde{X}_2|$, can be split along the lines of equation (17), and equations (18) and (19) then give a bound similar to that obtained in (7) in Section 3. Thus we have a bound on the objective function of the general marginals version of the LP((5)). Moreover, the bounds tend to zero as $n \to \infty$ as required.

These bounds help us prove results analogous to those in Section 4 for the case of marginals that have a density and a finite variance. We shall state these results here. The proofs need some additional technical arguments that are not of interest to this paper, and hence shall be omitted.

Theorem 9 *A covariance matrix is infeasible for the given marginals if, and only if, the chessboard LP is infeasible for some $n \geq 1$.*

Theorem 10 *Suppose that Σ is feasible. Then for all $\epsilon > 0$, there exists a chessboard distribution with covariance matrix Λ with the property that $r(\Sigma, \Lambda) < \epsilon$.*

We denote Ω to be the set of feasible covariance matrices, viewed as a subset of $d(d-1)/2$ dimensional Euclidean space.

Proposition 11 *The set Ω is nonempty, convex, closed and full-dimensional.*

Theorem 12 *There is a chessboard distribution of the form (16) with covariance matrix Σ if, and only if, $\Sigma \in \Omega^\circ$.*

We believe that the result in Theorem 7 on the size n of the linear programs holds for this general case too. The technical arguments involved in the proof seem more complicated, and is of current interest to us.

Acknowledgements

This research was partially supported by National Science Foundation Grants numbered DMI 0224884 and DMI 0230528.

REFERENCES

Devroye, L. 1986. Non-Uniform Random Variate Generation. New York: Springer-Verlag.

Ghosh, S., and S. G. Henderson. 2001. Chessboard distributions. In *Proceedings of the 2001 Winter Simulation Conference*, ed. B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer, 385–393. Piscataway NJ: IEEE.

Ghosh, S., and S. G. Henderson. 2002. Chessboard Distributions and Random Vectors with Specified Marginals and Covariance Matrix. *Operations Research* 50 (5): 820–834.

Johnson, M. E. 1987. Multivariate Statistical Simulation. New York: Wiley.

Johnson, N. L., and S. Kotz. 2004. Cloning of Distributions. *Preprint*.

Law, A. M., and W. D. Kelton. 2000. Simulation Modeling and Analysis. 3rd ed. New York: McGraw-Hill.

Mackenzie, G. R. 1994. Approximately Maximum-Entropy Multivariate Distributions with Specified Marginals and Pairwise Correlations. Ph. D. thesis, Department of Decision Sciences, University of Oregon, Eugene OR.

Nelsen, R. B. 1999. An Introduction to Copulas: Lecture Notes in Statistics, 139. New York: Springer-Verlag.

Sklar, A. 1959. Fonctions de répartition à n Dimensions et Leurs Marges. *Publications de l'Institut Statistique de l'Université de Paris* 8:229–231.

Walker, A. J. 1977. An Efficient Method for Generating Discrete Random Variables with General Distributions. *ACM Transactions on Mathematical Software* 3:253–256.